两种新的多维计算机化分类测验终止规则 *

任赫　　陈平

（北京师范大学中国基础教育质量监测协同创新中心，北京，100875）

摘 要 计算机化分类测验（*Computerized Classification Testing*, CCT）由于具备分类的功能，目前在职业资格考试、健康与护理问卷等以分类为目的的测验中得到广泛应用。作为 CCT 的重要组成部分，终止规则不仅决定测验停止的条件而且直接影响分类准确率及测验效率。然而，目前少有研究对多维 CCT（*Mulitidimensional CCT*, MCCT）的终止规则进行探索。针对已有 MCCT 终止规则的不足，提出两种新的 MCCT 终止规则（即基于马氏距离的多维序贯似然比规则 Mahalanobis-SPRT 和随机缩减的多维广义似然比规则 M-SCGLR），并开展模拟研究在不同实验条件下（比如，不同的题库结构、能力维度间相关及分界函数）考查它们的表现。结果表明：（1）在使用补偿性分界函数的条件下，Mahalanobis-SPRT 规则具有较高的分类精度和与同类方法相近的测验长度；（2）在几乎所有实验条件下，M-SCGLR 规则不仅在测验精度上大幅优于已有的多维随机缩减规则，而且具有较短的测验长度。

关键词 计算机化分类测验，终止规则，多维项目反应理论，马氏距离，随机缩减

1　引言

计算机化分类测验（*Computerized Classification Testing*, CCT）是一种特殊的计算机化自适应测验（*Computerized Adaptive Testing*, CAT），它能够高效地将被试划分到两个（即达标和未达标）或多个（比如，合格、良好和优秀）不同的类别中。CCT 将计算机作为测量媒介，使用自适应的选题策略和终止规则，基于被试当前的能力估计值精准地匹配测试题目，直到满足终止规则的要求，停止测验并提供对被试能力进行分类判断的结果。目前，这类测试已被广泛应用于职业资格考试（Huebner & Fina, 2015）和健康与护理问卷（Finkelman et al., 2011; Smits & Finkelman, 2013），其中的健康与护理问卷可以针对某种疾病或与护理计划直接相关的某些阶段将患者划分至有风险/无风险的类别中。尽管可以将各种心理测量理论作为 CCT

收稿日期：2020-06-04

* 国家自然科学基金面上项目（32071092）、中国基础教育质量监测协同创新中心基础教育质量监测科研基金项目（2019-01-082-BZK01 和 2019-01-082-BZK02）和中国基础教育质量监测协同创新中心自主课题（BJZK-2019A2-19003）资助。

通讯作者：陈平，Email：pchen@bnu.edu.cn

的基础，但是近年来大多数研究与应用都将焦点集中在基于项目反应理论（*Item Response Theory*, IRT）的可变长度的 CCT 研发上（Huebner & Fina, 2015; Li et al., 2020; Wang et al., 2020）。

完整的 CCT 和 CAT 均包括 IRT 模型、题库、选题策略、能力参数估计方法以及终止规则五个核心组成部分（郭磊 等, 2015）。但是两者在测验目的上并不相同：CAT 的目的是对被试能力进行准确估计（陈平, 2016），而 CCT 只需要输出对被试的类别划分。这就对测验的终止规则（也即测验应该如何停止以及如何给出测验结果）提出不同的要求，因此有必要对 CCT 的终止规则进行单独研究。在可变长度的二分类测验的背景下，已有的 CCT 终止规则可以被分为两类：似然比规则和贝叶斯规则。似然比规则的基本思路是通过事先规定不同类别被试的真实能力分界值，来构造似然比统计量并进行假设检验，从而完成对被试的分类。最早的似然比终止规则是 Wald（1947）提出的序贯似然比检验（*Sequential Probability Ratio Test*, SPRT）。Bartroff 等（2008）则将广义似然比（*Generalized Likelihood Ratio*, GLR）方法应用于变长的 CCT 中，并对 GLR 的良好性质进行证明。Thompson（2011）通过模拟研究发现：相比于 SPRT，GLR 方法能够在维持分类精度的基础上较大幅度地提高测验效率（或缩短测验长度）。此外，研究者发现：由于受制于现实因素（比如疲劳效应、练习效应），往往不可能要求被试一直作答直至满足 SPRT 的条件。在这种情况下，如果结合随机缩减（stochastic curtailment）技术就有可能提高测验效率。由此，Finkelman（2003, 2010）在 SPRT 的基础上结合随机缩减技术，开发出随机缩减的 SPRT（*Stochastically Curtailed SPRT*, SCSPRT）以及有预测能力的 SPRT（*SPRT with Predictive Power*, PPSPRT）。Huebner 和 Fina（2015）则将随机缩减技术与 GLR 方法相结合，提出基于 GLR 的随机缩减方法（*Stochastically Curtailed GLR*, SCGLR）。模拟研究的结果表明：使用随机缩减的方法能够提高测验效率（Finkelman, 2008; Huebner & Fina, 2015; Wang et al., 2020）。另一方面，贝叶斯规则的基本思路则是通过作答反应获取被试能力的后验分布，并使用后验分布计算损失函数值，从而完成对被试的分类。Lewis 和 Shehan（1990）率先引入先验函数和损失函数，并提出基于贝叶斯决策理论的终止规则。接下来，本文仅在变长的二分类测验情境下关注基于似然比规则的终止规则。

值得注意的是，上述终止规则都建立在单维 IRT（*Unidimensional IRT*, UIRT）的基础上，即假设测验仅考察被试单一维度的能力。但是在心理或教育测验的实践中，测验往往同时考察被试在多个维度上的潜在特质，这就与上述的单维性假设相悖（康春花, 辛涛, 2010）。为解决这一问题，基于多维 IRT（*Multidimensional IRT*, MIRT）构建多维 CCT（*Multidimensional CCT*, MCCT）就显得十分必要。迄今，关于 MCCT 的研究较少，只有少数研究者将特定的

似然比规则从单维情境推广至多维情境（Nydick, 2013）。在 MCCT 中，似然比规则的基本思路与单维情境的一致，但是能力参数的多维性导致各类别间的能力分界点转变为能力分界曲线（二维情境下）或能力分界曲面（三维及以上情境下）。为此，Nydick（2013）提出用似然函数约束的方法构建约束的 SPRT（*Constrained SPRT*, C-SPRT）、使用空间投影的方法构建投影的 SPRT（*Projected SPRT*, P-SPRT）以及在此基础上开发随机缩减的多维 SPRT（*Multidimensional SCSPRT*, M-SCSPRT）。此外，Nydick（2013）还首先将多维 GLR 检验（*Multidimensional GLR*, M-GLR）引入 MCCT。

综上，基于 MIRT 构建 MCCT 终止规则能够更好地适应现实测验的需要。本文在总结与分析已有 MCCT 终止规则的基础上，提出两种新的 MCCT 终止规则：第一种是基于马氏距离的多维序贯似然比终止规则（*Mahalanobis-SPRT*），具体思路是将马氏距离融入 P-SPRT 方法；第二种是多维随机缩减的 GLR 规则（*Multidimensional SCGLR*, M-SCGLR），可以被视为 SCGLR 在多维情境下的推广。两种新终止规则相对于已有规则的表现，将通过模拟研究在多种实验条件下进行全面评价。

本文的剩余部分将按如下方式进行组织：第 2 节首先简要描述本文使用的 MIRT 模型以及四种已有的 MCCT 终止规则（即 C-SPRT、P-SPRT、M-GLR 以及 M-SCSPRT），然后详细介绍两种新提出的 MCCT 终止规则（即 Mahalanobis-SPRT 和 M-SCGLR）。第 3 节将介绍模拟研究设计，并在第 4 节展示研究结果与结论。最后一节进行讨论并展望未来的研究方向。

2 方法

2.1 MIRT 模型

本文假设所有题目都由多维三参数逻辑斯蒂克模型（*Multidimensional Three-Parameter Logistic Model*, M3PL）建模。在该模型中，能力向量为$\boldsymbol{\theta}_i$的被试i正确作答二级计分题目j的概率为（Reckase & Mckinley, 1982），

$$P_j(\boldsymbol{\theta}_i) \equiv \text{Prob}\left(Y_{ij} = 1 \middle| \boldsymbol{\theta}_i, \mathbf{a}_j, d_j, c_j\right) = c_j + \frac{1-c_j}{1+\exp\left[-\left(\mathbf{a}_j^T\boldsymbol{\theta}_i + d_j\right)\right]}, \quad (1)$$

其中，Y_{ij}是取值为 0 或 1 的伯努利随机变量，表示被试i在题目j上的二级计分作答反应。$\boldsymbol{\theta}_i = \left(\theta_{i1}, \theta_{i2}, \dots, \theta_{ip}\right)^T$表示被试$i$的$p$维能力向量，$T$表示转置。$\mathbf{a}_j = \left(a_{j1}, a_{j2}, \dots, a_{jp}\right)^T$为题目$j$的$p$维区分度参数向量，有$\mathbf{a}_j^T\boldsymbol{\theta}_i = \sum_{k=1}^{p} a_{jk}\,\theta_{ik}$；标量$d_j$是与题目难度相关的截距参数，标量$c_j$则是题目的伪猜测参数。为方便对模型参数的含义进行解释，Ackerman（1994）定义$MDISC_j = \left(a_{j1}^2 + a_{j2}^2 + \dots + a_{jp}^2\right)^{1/2}$作为题目$j$的多维区分度（multidimensional discrimination），

定义$MDIFF_j = \frac{-d_j}{MDISC_j}$作为题目$j$的多维难度（multidimensional difficulty）。

由此，被试i对j'道题目的作答反应$\mathbf{Y}_{ij'} = (Y_{i1}, Y_{i2}, \ldots, Y_{ij'})$的似然函数为，

$$L(\boldsymbol{\theta}_i|\mathbf{Y}_{ij'}) = \prod_{j=1}^{j'}[P_j(\boldsymbol{\theta}_i)]^{Y_{ij}}[Q_j(\boldsymbol{\theta}_i)]^{1-Y_{ij}}, \quad (2)$$

其中，$Q_j(\boldsymbol{\theta}_i) = 1 - P_j(\boldsymbol{\theta}_i)$，表示被试$i$错误作答题目$j$的概率。

2.2 已有的 MCCT 终止规则

2.2.1 似然比思想与 C-SPRT、P-SPRT 以及 M-GLR 规则

目前大多数关于 MCCT 的终止规则都是基于单维 CCT 的终止规则而构建。一个单维序贯似然比规则的构成可以总结为四个步骤：（1）构造假设检验；（2）确定不同等级间的能力阈值θ_0；（3）在θ_0处给定一个δ邻域，即$(\theta_0 - \delta, \theta_0 + \delta) \equiv (\theta_l, \theta_u)$。当能力值落在该区间时，认为未获得足够信息对被试进行分类，因此该区间也被称为无差别区间；而当能力值大于θ_u时认为被试属于“达标”的类别，当能力值小于θ_l时认为被试属于“未达标”的类别；（4）构建似然比统计量并确定拒绝域。在将终止规则从 CCT 推广到 MCCT 时，需要定义能力分界曲线或曲面才能将不同类别的被试区分开来。由此，单维情境下的能力分界点θ_0就变为多维空间中的能力分界曲线或曲面$g(\boldsymbol{\theta}) = 0$，其中$g(\boldsymbol{\theta})$为分界函数，具体可分为补偿性的分界函数（比如，$g(\boldsymbol{\theta}) = \theta_1 + \theta_2$）和非补偿的分界函数（比如，$g(\boldsymbol{\theta}) = \begin{cases}\theta_1, 若\theta_2 \geq 0 \\ \theta_2, 若\theta_1 > 0\end{cases}$）。此时的一个研究问题是如何将分界曲线或曲面转化为单维情况下的分界阈值θ_0。另外，即使获得θ_0，多维空间中的θ_0在不同方向上可以构造任意多个δ邻域，因此如何选择θ_l和θ_u是另一个需要解决的问题。C-SPRT、P-SPRT 以及 M-GLR 分别从三个不同的角度提供解决方案。

首先，似然比规则需要构造假设检验，

$$\begin{aligned} H_0&: \boldsymbol{\theta} \in \boldsymbol{\Theta}_n \\ H_1&: \boldsymbol{\theta} \in \boldsymbol{\Theta}_m, \end{aligned} \quad (3)$$

其中，$\boldsymbol{\Theta}_m$表示属于“达标”类别的被试的能力空间，$\boldsymbol{\Theta}_n$表示属于“未达标”类别的被试的能力空间。于是，接受原假设H_0表示被试属于“未达标”类别，接受备择假设H_1则表示被试属于“达标”类别。

（1）C-SPRT

基于构造的上述假设，C-SPRT 的基本思路是使用约束在分界曲线或曲面上的能力估计值替代能力分界点θ_0，并计算相应的$\boldsymbol{\theta}_l$和$\boldsymbol{\theta}_u$（分别对应$\boldsymbol{\Theta}_n$的上界与$\boldsymbol{\Theta}_m$的下界上的点）。具体地说，在被试i作答完j'道题目后，C-SPRT 方法首先将在分界曲线或曲面上计算得到的能力参数估计值$\hat{\boldsymbol{\theta}}_0$作为能力分界点$\theta_0$的估计，即

$$\widehat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\arg\max}\left[\log L\left(\boldsymbol{\theta}\middle|\mathbf{Y}_{ij'}\right)\right], \tag{4}$$

其中，$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}: g(\boldsymbol{\theta}) = 0\}$表示能力分界曲线或曲面。上式表示将$\boldsymbol{\Theta}_0$上使得$\log L\left(\boldsymbol{\theta}\middle|\mathbf{Y}_{ij'}\right)$取最大值的点记为$\widehat{\boldsymbol{\theta}}_0$。C-SPRT 方法然后在$\widehat{\boldsymbol{\theta}}_0$处$g(\boldsymbol{\theta}) = 0$的法向量方向上构造$\delta$邻域。记$\boldsymbol{\theta}_\delta$为该方向的单位向量，即$\boldsymbol{\theta}_\delta = \frac{\nabla g(\widehat{\boldsymbol{\theta}}_0)}{\|\nabla g(\widehat{\boldsymbol{\theta}}_0)\|_2}$，其中$\nabla$为哈密顿算子，表示微分运算，$\|\cdot\|_2$表示欧几里得范数，用于衡量欧氏空间内的距离。于是可得到无差别区间的上下限分别为

$$\widehat{\boldsymbol{\theta}}_u = \widehat{\boldsymbol{\theta}}_0 + \delta\boldsymbol{\theta}_\delta, \tag{5}$$

$$\widehat{\boldsymbol{\theta}}_l = \widehat{\boldsymbol{\theta}}_0 - \delta\boldsymbol{\theta}_\delta. \tag{6}$$

根据 Wald（1947）提出的似然比检验构造似然比统计量，得到

$$C_{ij'} = \log\left[\mathrm{LR}\left(\widehat{\boldsymbol{\theta}}_u, \widehat{\boldsymbol{\theta}}_l\middle|\mathbf{Y}_{ij'}\right)\right] = \log\left[\frac{L\left(\widehat{\boldsymbol{\theta}}_u\middle|\mathbf{Y}_{ij'}\right)}{L\left(\widehat{\boldsymbol{\theta}}_l\middle|\mathbf{Y}_{ij'}\right)}\right]. \tag{7}$$

记第一类和第二类错误率分别为α和β，令$A = A(\alpha,\beta)$、$B = B(\alpha,\beta)$、$C_l = \log(A)$、$C_u = \log(B)$且$C_0 = \frac{C_l + C_u}{2}$。在分类测验的背景下，通常取$A(\alpha,\beta) = \frac{\beta}{1-\alpha}$和$B(\alpha,\beta) = \frac{1-\beta}{\alpha}$（Finkelman, 2003）。在被试$i$作答完$j'$道题目后，计算$C_{ij'}$，并基于似然比检验规则给出如下判断：若

$$C_{ij'} \le C_l, \tag{8}$$

则停止测验，测验长度为j'，并判断被试属于“未达标”；若

$$C_{ij'} \ge C_u, \tag{9}$$

则停止测验，测验长度为j'，并判断被试属于“达标”；否则，即

$$C_l < C_{ij'} < C_u, \tag{10}$$

则继续给被试作答下一道题。

Wald-Wolfowitz 定理表明：在测验可以持续进行直至满足上述终止规则的情况下，SPRT 是具有同等检验力的检验中所需观测个数最少的假设检验，即最优序贯检验（Wald & Wolfowitz, 1948）。但是在现实情境下，由于疲劳效应、练习效应等因素的影响，不可能要求被试一直作答直至满足不等式（8）或（9）。因此在单维 CCT 中，一般通过事先设定最大测验长度J以满足上述现实需要。于是，研究者在设计不定长的 MCCT 终止规则时也沿用这一附加的强制结束条件。这就是说，若达到最大长度J时测验仍未结束，则根据下述准则对被试进行强制分类：若$C_{iJ} \le C_0$，则停止测验，测验长度为J，并判断被试属于“未达标”；若$C_{iJ} > C_0$，则停止测验，测验长度为J，并判断被试属于“达标”（记该准则为最大测验长度下似然比检验的判断准则）。

记被试最终完成测验的实际作答题目数为K，分类判断结果为D（$D = m$表示被试属于“达标”，$D = n$表示被试属于“未达标”），最大测验长度为J，则整个 C-SPRT 的判断规则

可以概括如下，

$$\begin{cases}\text{停止测验}, K=j', D=n & \text{若}\{j'<J, C_{ij'}\le C_l\}\text{或}\{j'=J, C_{ij'}\le C_0\}\\ \text{停止测验}, K=j', D=m & \text{若}\{j'<J, C_{ij'}\ge C_u\}\text{或}\{j'=J, C_{ij'}> C_0\}.\\ \text{继续测验} & \text{否则}\end{cases} \quad (11)$$

（2）P-SPRT

P-SPRT 与 C-SPRT 唯一的区别在于它采用不同方法将分界曲线或曲面转换为可用于假设检验的分界点。具体地说，P-SPRT 将基于当前作答得到的被试能力估计值投影至$g(\boldsymbol{\theta})=0$所刻画的边界上，并将投影视作分界点。在被试作答完j'道题目后，对其能力估计值进行投影的表述如下：

$$\widehat{\boldsymbol{\theta}}_0=\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\operatorname{argmin}}\left\|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}\right\|_2, \quad (12)$$

其中，$\widehat{\boldsymbol{\theta}}_i$表示被试$i$的能力估计值。由于$\|\cdot\|_2$代表欧氏空间内的距离，因此公式（12）表示将$\boldsymbol{\Theta}_0$上与$\widehat{\boldsymbol{\theta}}_i$距离最近的点记为$\widehat{\boldsymbol{\theta}}_0$，也就是将$\widehat{\boldsymbol{\theta}}_i$投影至$\boldsymbol{\Theta}_0$上，并将投影得到的点记为$\widehat{\boldsymbol{\theta}}_0$。确定$\widehat{\boldsymbol{\theta}}_0$后，P-SPRT 也依照等式（5）、（6）和（7）得到$\widehat{\boldsymbol{\theta}}_u$、$\widehat{\boldsymbol{\theta}}_l$以及$C_{ij'}$。

（3）M-GLR

M-GLR 方法在构造似然比统计量的思路上与 P-SPRT 和 C-SPRT 都不同。GLR 统计量$C_{ij'}$是似然函数在不同类别被试（即“达标”与“未达标”）的能力空间中的最大值之比的对数，它不同于等式（7）需要确定$\widehat{\boldsymbol{\theta}}_u$和$\widehat{\boldsymbol{\theta}}_l$，因此从理论上可避免“多维情境下要将分界曲线或曲面转换为分界点”的需求。M-GLR 统计量定义为

$$C_{ij'}=\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}\left[L\left(\boldsymbol{\theta}_1|\mathbf{Y}_{ij'}\right)\right]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}\left[L\left(\boldsymbol{\theta}_2|\mathbf{Y}_{ij'}\right)\right]}. \quad (13)$$

其中，$\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m$表示$\boldsymbol{\theta}_1$是“达标”被试的能力空间$\boldsymbol{\Theta}_m$中的任一值，$\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n$表示$\boldsymbol{\theta}_2$为“未达标”被试的能力空间$\boldsymbol{\Theta}_n$中的任一值。公式（13）的分子部分即为在“达标”被试的能力空间内似然函数的最大值，而分母部分即为在“未达标”被试的能力空间内似然函数的最大值。由此可以发现，与单维的 GLR 方法相比，M-GLR 只是将$C_{ij'}$中求极值的集合由单维的能力区间变为多维能力空间中的区域，其性质并没有变化。需要说明的是，尽管 C-SPRT、P-SPRT 与 M-GLR 在构造统计量的具体方式上存在差异，但都是基于公式（3）所对应的假设检验，而且仅依赖构造的似然比统计量进行判断。因此，在得到$C_{ij'}$（序贯似然比统计量或广义似然比统计量）后，P-SPRT 和 M-GLR 的判断准则也都通过与C_l、C_u或C_0进行比较得到测验结果，即按照公式（11）所定义的规则对被试做出分类判断。

2.2.2 随机缩减与 M-SCSPRT 规则

如前所述，由于最大测验长度J的引入与 Wald-Wolfowitz 定理的前提假定相悖，因此在同等条件下，SPRT 不再具有最大检验力。这种低效不仅增大测验长度，而且导致测验时间和题目曝光率的上升。因此，在维持 SPRT 分类准确率的基础上缩短测验长度有助于 MCCT 的应用。随机缩减（Finkelman, 2008; Huebner & Fina, 2015）正是解决该问题的一种方法：即如果被试接下来的作答反应在较大概率上不会改变当前对被试的分类判断，那么此时便结束测验是合理的。

M-SCSPRT 规则是一种将随机缩减与 C-SPRT 相结合的多维似然比终止规则，它在完整保留公式（11）所定义判断准则的基础上，对原本需要继续作答的被试i再次进行判断。具体地说，M-SCSPRT 按照等式（4）至等式（7）计算约束下的似然比统计量$C_{ij'}$，并计算被试作答至最大测验长度J时，对被试的分类判断与当前一致的概率

$$P\left(D_J = D_{j'} \middle| C_{ij'}\right), \tag{14}$$

其中$D_{j'}$表示被试作答完j'道题目时，对被试的预分类；D_J表示被试作答完J道题目时，对被试的最终分类。预分类的判断准则与最大测验长度下似然比检验的判断准则一致，即

$$\begin{cases} D_{j'} = n, & C_{ij'} \le C_0 \\ D_{j'} = m, & C_{ij'} > C_0 \end{cases}. \tag{15}$$

由公式（11），在 2.2.1 节所述的三种似然比检验中，若$j' < J$且$C_l < C_{ij'} < C_u$，测验将继续进行。但是，M-SCSPRT 方法在$j' < J$时，对公式（11）进行了如下调整，而在$j' = J$时不变，

$$\begin{cases} \text{停止测验}, K = j', D = n & \text{若}\{C_{ij'} \le C_l\}\text{或}\{C_l < C_{ij'} \le C_0, P(D_J = n|C_{ij'}) \ge 1 - \epsilon_1\} \\ \text{停止测验}, K = j', D = m & \text{若}\{C_{ij'} \ge C_u\}\text{或}\{C_u > C_{ij'} > C_0, P(D_J = m|C_{ij'}) \ge 1 - \epsilon_2\}, \\ \text{继续测验} & \text{否则} \end{cases} \tag{16}$$

其中，ϵ_1与ϵ_2为事先设定的临界值。以往的模拟研究表明：当ϵ_1与ϵ_2都取 0.05 时，能在损失较小测验分类精度的前提下大幅缩短测验长度（Finkelman, 2008, 2010）。

2.3 两种新的 MCCT 终止规则

2.3.1 Mahalanobis-SPRT

如公式（12）所述，P-SPRT 规则使用欧氏距离对被试能力估计值进行空间投影。但是，在 CCT 施测的初期阶段，对被试能力的估计往往不够准确，P-SPRT 仅使用一次估计的能力结果$\hat{\boldsymbol{\theta}}_i$进行投影可能会使$\hat{\boldsymbol{\theta}}_0$不够稳定，从而影响分类结果（在高维情境下这种影响可能会尤为突出）。因此在多维情境中，“按欧氏距离对被试能力值进行投影”这种做法有待商榷。

本文基于聚类分析的思想，提出基于马氏距离的 Mahalanobis-SPRT 规则，以克服 P-SPRT 的上述不足。在测验初期，尽管单个被试能力估计值并不准确，但是如果将多个能力估计值

综合起来，就可以大致描绘出被试真实能力值所处的范围。具体地说，测验初期的能力估计值是在真值附近上下波动的，而并非一致地高于或低于真值，所以多个能力估计值的均值，往往就更加接近真值。图 1 表示某名被试在一个二维测验过程中，其能力估计值随作答题目数量变化而变化的情况。其中，蓝色的三角形点代表该被试的能力真值，红色的圆形点代表对被试能力的估计值，红色越深表示得到该能力估计值时被试作答的题目数量越多。由图 1 可以看到：在被试作答的题目数量较少时，被试的能力估计值与真值相差较大。但与此同时，在两个维度上，能力估计值都是围绕真值上下波动的。因此，在测验初期，多个能力估计值的均值就能够对被试能力真值进行比较准确的描述。

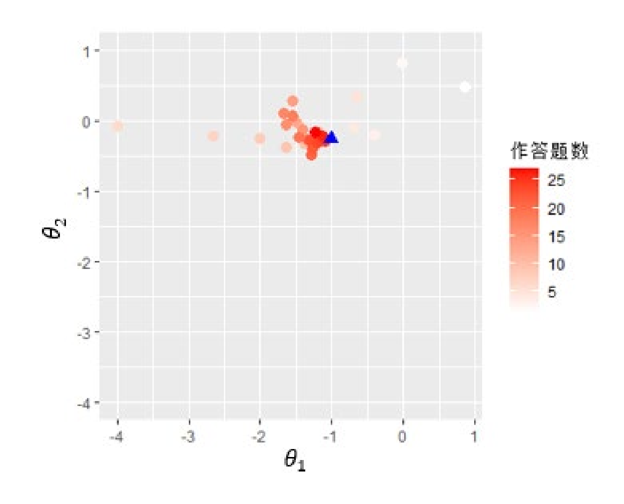


图 1 二维情境下某名被试的能力估计值随作答题数的变化图

综上，使用分界曲线或曲面上的点中，到“已得到的多个能力估计值的均值”的马氏距离最近的点作为$\widehat{\boldsymbol{\theta}}_0$（这也正是 Mahalanobis-SPRT 方法的做法）比 P-SPRT 中直接使用分界曲线或曲面上到$\widehat{\boldsymbol{\theta}}_i$的欧氏距离最近的点要更合理。于是，我们可以定义 Mahalanobis-SPRT 规则下的分界点$\widehat{\boldsymbol{\theta}}_0$，即

$$\widehat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\operatorname{argmin}}\left\|\overline{\boldsymbol{\theta}}_{ij'} - \boldsymbol{\theta}\right\|_M. \quad (17)$$

其中，$\|\cdot\|_M$代表马氏距离；$\overline{\boldsymbol{\theta}}_{ij'}$是被试$i$作答完$j'$道题目后得到的$j'$个能力估计值的均值，代表对被试能力真值的近似刻画。如果将被试i作答完第j道题目后得到的p维能力估计值记为$\widehat{\boldsymbol{\theta}}_{ij} = (\hat{\theta}_{ij1}, \hat{\theta}_{ij2}, \dots, \hat{\theta}_{ijp})$，那么$\overline{\boldsymbol{\theta}}_{ij'} = \left(\sum_{j=1}^{j'}\hat{\theta}_{ij1}/j', \sum_{j=1}^{j'}\hat{\theta}_{ij2}/j', \dots, \sum_{j=1}^{j'}\hat{\theta}_{ijp}/j'\right)$。另外，如果将这$j'$个能力估计值的协方差矩阵记为$\boldsymbol{\Sigma}_{ij'}$，那么根据马氏距离的定义且当$\boldsymbol{\Sigma}_{ij'}$可逆（即$\left|\boldsymbol{\Sigma}_{ij'}\right| \neq 0$）时，有

$$\widehat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\operatorname{argmin}}\left\|\overline{\boldsymbol{\theta}}_{ij'} - \boldsymbol{\theta}\right\|_M = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}{\arg\min}\sqrt{(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_{ij'})\boldsymbol{\Sigma}_{ij'}^{-1}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_{ij'})^{\mathrm{T}}}. \quad (18)$$

定义$\boldsymbol{\theta}_\delta$为$\bar{\boldsymbol{\theta}}_{ij'}$与$\widehat{\boldsymbol{\theta}}_{\mathbf{0}}$的差向量方向上的单位向量，即

$$\boldsymbol{\theta}_\delta = \frac{\bar{\boldsymbol{\theta}}_{ij'} - \widehat{\boldsymbol{\theta}}_0}{\|\bar{\boldsymbol{\theta}}_{ij'} - \widehat{\boldsymbol{\theta}}_0\|_2}. \tag{19}$$

确定$\widehat{\boldsymbol{\theta}}_0$与$\boldsymbol{\theta}_\delta$后，Mahalanobis-SPRT 按照等式（5）、（6）和（7）得到$\widehat{\boldsymbol{\theta}}_u$、$\widehat{\boldsymbol{\theta}}_l$以及$C_{ij'}$，然后按照公式（11）所述的判断准则对被试进行分类。

需要指出的是，公式(17)中有两处与 P-SPRT 所定义的公式(12)不同：第一，Mahalanobis-SPRT 使用"已得到的多个能力估计值的均值（$\bar{\boldsymbol{\theta}}_{ij'}$）"代替 P-SPRT 中的单个能力估计值；第二，Mahalanobis-SPRT 使用马氏距离作为距离的度量方式，而非 P-SPRT 中的欧式距离。总之，Mahalanobis-SPRT 新规则使用被试能力的一系列序贯估计值，并将分界曲线或曲面上距其均值最近的点作为$\widehat{\boldsymbol{\theta}}_0$；相较于 P-SPRT 使用单一能力估计值进行投影，新规则理应能够获得更加稳健的结果。

2.3.2 M-SCGLR 规则

Huebner 和 Fina（2015）提出的 SCGLR 规则将随机缩减的方法与 GLR 相结合，在保持测验分类精度的前提下能够缩短测验长度。因此，本文将 SCGLR 方法推广至 MCCT 情境，并得到多维的 SCGLR 规则(记为 M-SCGLR)。M-SCGLR 直接沿用 M-GLR 的方式构造 GLR 统计量，如等式（13）所示。

但是，在对被试进行分类判断时，M-SCGLR 采用的是随机缩减方法，即对公式（11）在$j' < J$的情况下进行如下调整：

$$\begin{cases} \text{停止测验}, K = j', D = n \quad \text{若}\{C_{ij'} \le C_l\}\text{或}\{C_l < C_{ij'} \le C_0, P(D_J = n|C_{ij'}) \ge 1 - \epsilon_1\} \\ \text{停止测验}, K = j', D = m \;\text{若}\{C_{ij'} \ge C_u\}\text{或}\{C_u > C_{ij'} > C_0, P(D_J = m|C_{ij'}) \ge 1 - \epsilon_2\}, \\ \qquad\qquad \text{继续测验} \qquad\qquad\qquad\qquad\qquad\qquad \text{否则} \end{cases} \tag{20}$$

在公式（20）中，

$$P(D_J = n|C_{ij'}) = 1 - P(D_J = m|C_{ij'}) \approx \Phi\left(\frac{C_0 - \mathbb{E}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})}{\sqrt{\mathrm{Var}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})}}\right), \tag{21}$$

其中，

$$\mathbb{E}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'}) = C_{ij'} + \textstyle\sum_{j=j'+1}^{J} \mathbb{E}_{\boldsymbol{\theta}}\left(\log \frac{\sup_{\boldsymbol{\theta}_1 \in \Theta_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2 \in \Theta_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}\right), \tag{22}$$

$$\mathrm{Var}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'}) = \textstyle\sum_{j=j'+1}^{J} \mathrm{Var}_{\boldsymbol{\theta}}\left(\log \frac{\sup_{\boldsymbol{\theta}_1 \in \Theta_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2 \in \Theta_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}\right), \tag{23}$$

其中，$\Phi(\cdot)$为标准正态分布的分布函数。公式（21）、（22）和（23）的具体推导过程，感兴趣的读者可参见附录 1。

需要注意的是，当能够事先知道第$j'+1$道题到第J道题目的选取时，上述计算并不困难。但在自适应序贯选题的情境下，无法提前获知接下来的题目。此时，为计算等式(22)与(23)，可以使用一组合适的题目替代被试接下来要实际作答的题目。比如，当使用 D 最优选题策略时（即最大化 Fisher 信息矩阵的行列式），可以基于当前的能力估计值计算所有剩余题目的 Fisher 信息矩阵的行列式，然后选择值最大的$J-j'$道题目作为替代的题目。单维情形下的研究表明：当使用替代题目时，需适当减小错误率ϵ_1和ϵ_2的值（Finkelman, 2008）。

3　实验

本研究采用 R 3.4.2 自编计算机程序开展模拟研究，共有两个研究目的：（1）将新提出的 2 种 MCCT 终止规则（即 Mahalanobis-SPRT 和 M-SCGLR）与已有的 4 种终止规则（即 C-SPRT、P-SPRT、M-GLR 以及 M-SCSPRT）进行比较，并评价它们在测验精度和测验效率两方面的表现以揭示各种方法的优缺点以及适用情境；（2）对比上述 6 种终止规则对于具有特定能力水平的被试的分类表现，以探究各种规则对特定被试的分类敏感度是否存在明显差异。

考虑到不同题库结构、能力维度间相关及能力分界曲线会对 MCCT 的结果产生影响，本研究设置 2 种题库结构、3 种能力维度间的相关水平和 2 种分界曲线，对 6 种 MCCT 终止规则展开模拟研究，也即采用 2×3×2×6 的实验设计（共产生 72 种实验条件，12 种 MCCT 测验情境）实现研究目的一。另外，本研究还分别选取靠近或远离能力分类曲线的 36 种特定能力取值的被试及 2 种分界曲线以实现研究目的二。

3.1　题库与被试生成

MIRT 的研究中通常考虑两种题库结构，即题目内多维（*within-item multidimensionality*）和题目间多维（*between-item multidimensionality*）。其中，题目内多维是指题库中的每道题目均测量一个或多个维度，而题目间多维则是指题库中的每道题目有且仅测量一个维度（Hartig & Höhler, 2008; Wang & Chen, 2004）。由于题库结构会对被试能力向量的估计精度产生影响（Chen & Wang, 2016），因此本文按照公式（1）所定义的 MIRT 模型生成两个 MCCT 题库：题库 1 采用题目内多维的结构，题库 2 采用题目间多维的结构，每个题库均包含 900 道题目。题库 1 中每道题目都测量两个维度（即$p=2$），由此可记题库中的题目参数向量为$\boldsymbol{\gamma}=(a_1,a_2,d,c)^T$。题库 2 中的一半题目仅测量第一个维度，另一半则仅测量第二个维度。为使模拟情境尽可能地接近现实情况，本研究按照 Nydick（2013）的做法模拟各个参数：

（1）a_1和a_2参数。从均值为$\mu_{log}=0.5$、标准差为$\sigma_{log}=0.1$的对数正态分布中随机抽取

区分度参数$MDISC$（即$\sqrt{a_1^2+a_2^2}$）。于是在题库 1 中，可以从均匀分布$U(0, MDISC^2)$中抽取a_1^2，则$a_2^2 = MDISC^2 - a_1^2$；而在题库 2 中，设置对应维度的a参数等于$MDISC$即可。

（2）d和c参数。从$U(-3.5, 3.5)$中随机抽取与难度相关的参数b。由此，在题库 1 中，与之相关的参数$d = -b \cdot \mathbf{a1}$，其中$\mathbf{1}$是元素全为 1 的 2 维列向量；在题库 2 中，d为b和对应维度的a参数的乘积的负值。此外，固定参数c为 0.2。

另一方面，本文模拟 3000 名被试参与测验，被试的能力向量$\boldsymbol{\theta} = (\theta_1, \theta_2)^T$随机抽取自均值向量为$\boldsymbol{\mu} = (0, 0)$、协方差矩阵为$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$的二维正态分布 $\mathbf{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，其中$\rho = 0$、0.5 和 0.8，分别对应能力维度间没有相关、中等相关和高度相关 3 种水平（Chen et al., 2017）。此外，本文模拟 36 个特定的能力向量值(θ_1, θ_2)用于实现研究目的二，其中$\theta_1, \theta_2 \in \{-0.5, -0.3, -0.1, 0.1, 0.3, 0.5\}$（6 个点在两个维度上完全交叉共形成 36 个点），每个能力点上生成 500 名被试参与测验。

对模拟生成的数据进行描述统计，得到的结果如表 1 所示。

表 1 研究 1 中各参数的描述统计表

	题库 1（题目内多维）				题库 2（题目间多维）				被试（$\boldsymbol{\rho} = \mathbf{0}$）		被试（$\boldsymbol{\rho} = \mathbf{0.5}$）		被试（$\boldsymbol{\rho} = \mathbf{0.8}$）	
统计量	a_1	a_2	d	c	a_1	a_2	d	c	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
平均数	1.103	1.098	0.086	0.200	0.830	0.833	0.131	0.200	−0.010	0.021	0.022	0.006	−0.016	−0.025
标准差	0.428	0.414	4.348	0.000	0.839	0.842	3.336	0.000	0.998	0.996	1.011	0.991	0.999	1.000
最小值	0.038	0.040	−9.327	0.200	0.000	0.000	−6.281	0.200	−3.331	−3.125	−3.614	−3.196	−4.016	−3.267
最大值	2.285	2.065	8.873	0.200	2.196	2.329	7.220	0.200	3.252	3.332	4.269	3.071	3.264	3.712
相关系数矩阵	1	−0.782	−0.011	—	1	−0.981	−0.001	—	1	−0.002	1	0.486	1	0.803
	0.782	1	0.009	—	−0.981	1	0.004	—	−0.002	1	0.486	1	0.803	1
	−0.011	0.009	1	—	−0.001	0.004	1	—						

注：表 1 中仅呈现为实现研究目的一而生成的各参数的描述统计量，这是因为研究目的二是针对 36 种特定能力值的被试。

3.2 MCCT 的模拟程序描述

从能力估计方法、选题策略以及终止规则等三个方面对 MCCT 的模拟过程进行描述：

（1）能力估计方法

本研究采用约束的极大似然估计法（*Maximum Likelihood Estimation*, MLE）估计被试的能力向量值$\boldsymbol{\theta}$，记为$\hat{\boldsymbol{\theta}}$，参数的估计范围限定在$[-4, 4] \times [-4, 4]$的正方形区域，公式如下，

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in [-4,4] \times [-4,4]}{\arg\max} \{\log[L(\boldsymbol{\theta}|\mathbf{Y})]\}. \tag{24}$$

具体的估计过程由 R 3.4.2 中的 donlp2 函数实现。

（2）选题策略

根据以往研究（Nydick, 2013; Segall, 1996），使用经典的 D 最优（D-optimality）策略选取题目。D 最优策略选择最大化 Fisher 信息矩阵的行列式的题目，它等价于选择最小化未知参数协方差矩阵的行列式（即$\boldsymbol{\theta}$的置信椭球体积）的题目。针对（1）式定义的 MIRT 模型，任一题目j的 Fisher 信息矩阵为，

$$I_j(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \log[L(\boldsymbol{\theta}|Y)]}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right] = \frac{[1-p_j(\boldsymbol{\theta})][p_j(\boldsymbol{\theta})-c_j]^2}{p_j(\boldsymbol{\theta})[1-c_j]^2}\mathbf{a}_j\mathbf{a}_j^{\mathrm{T}}. \tag{25}$$

在模拟的二维情境下，上述 Fisher 信息矩阵是一个2×2的矩阵。在选择第j道题目时，关于$\boldsymbol{\theta}$的 Fisher 信息矩阵是“已作答的$j-1$道题目的信息矩阵”与“候选的第j道题的信息矩阵”之和（也即$\sum_{k=1}^{j-1} I_k(\boldsymbol{\theta}) + I_j(\boldsymbol{\theta})$）。由此，被试能力估计值的置信椭球的体积为，

$$\mathrm{Var}\ (\boldsymbol{\theta}) = \left|\sum_{k=1}^{j-1} I_k(\boldsymbol{\theta}) + I_j(\boldsymbol{\theta})\right|^{-1}, \tag{26}$$

使用 D 最优策略选择的第j道题目就是剩余题库中使得公式（26）达到最小的题目。

另外，由于在测验初期难以获得对被试能力的准确估计值（Chang & Ying, 1996），难以达到精准选题的目的。因此，本研究中每次测验的前 4 道试题从题库中随机抽取产生。

（3）终止规则与分界曲线

本研究采用 C-SPRT、P-SPRT、M-GLR、M-SCSPRT、Mahalanobis-SPRT、M-SCGLR 等 6 种规则终止测验。按照 Thompson（2010）的设置，这里令$\alpha = \beta = 0.1$且$\epsilon_1 = \epsilon_2 = 0.025$。为考察不同类型的分界曲线对结果的影响，本研究设置两种分界曲线：补偿性分界曲线和非补偿的分界曲线。其中，补偿性分界曲线是指不同维度间的能力是通过线性组合的方式结合在一起，此时被试在某个维度上能力的不足可以由其在其他维度上的高能力来补偿。否则，当被试不同维度间的能力无法相互补偿时，即为非补偿性分界曲线。参考 Nydick（2013）的做法，本研究选取的补偿性分界曲线为$g(\boldsymbol{\theta}) = \theta_1 + \theta_2 = 0$，非补偿的分界曲线为$g(\boldsymbol{\theta}) = \begin{cases}\theta_1 = 0, \theta_2 \geq 0\\ \theta_2 = 0, \theta_1 > 0\end{cases}$（也即笛卡尔坐标系中构成第一象限的坐标轴）。

3.3 评价指标

选择平均测验长度（*Average Test Length*, ATL）、正确分类率（*Percent of Correct Classification*, PCC）以及损失函数（*loss*）评价每种终止规则。

ATL 是某种测验情境下所有被试的最终测验长度的平均值，在一定程度上反映测验效率。PCC 是被正确分类的被试占该测验情境下所有被试的比例，反映测验分类精度。Finkelman（2010）定义的 loss 是对某次测验的测验精度和效率的综合评价指标，

$$\mathrm{loss} = R \times \mathbf{1}_W + K, \tag{27}$$

其中，$\mathbf{1}_W$表示错误分类的示性函数（当对被试错误分类时取值为 1，当没有误判时取值为 0）；R表示错误分类的惩罚程度，一般为非负值，其取值越大就表示对错误分类的惩罚越大，即对精度的要求越高（在计算时需由研究者根据测验对错误分类的厌恶程度给定）；K和前文一样，表示被试最终完成测验时实际作答的题目数。在某次测验中，如果将被试错误分类，loss 的值为惩罚值R与该次测验的长度之和；否则，loss 的值就等于该次测验的长度。如果在多次测验中固定R，并将式（27）取平均，即可得到平均损失，

$$\overline{\text{loss}} = R \times (1 - \text{PCC}) + \text{ATL}, \quad (28)$$

由此，平均损失是一种结合 PCC 和 ATL 的综合评价指标。具体地说，R确定后，对于某个终止规则而言，其 PCC 越大且 ATL 越小，则平均损失就越小，表示该方法表现越好；相反地，PCC 越小，ATL 越大，平均损失就越大，表示该方法表现越差。根据平均损失的大小，就可以指导实际测验中终止规则的选择。

4 结果

4.1 各种规则的分类精度与效率

图 2 呈现了 6 种终止规则在各种 MCCT 测验情境下的 ATL 及 PCC 结果。

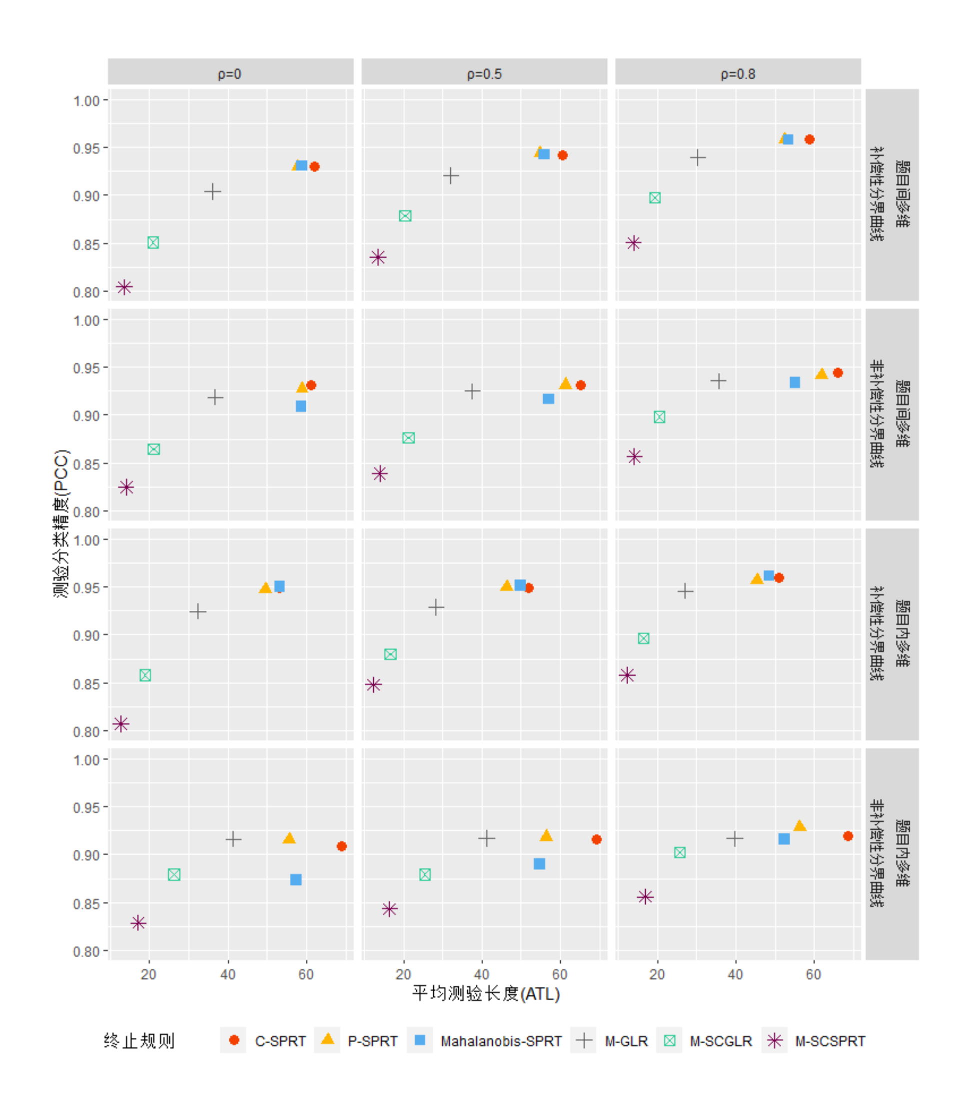


图 2 6 种终止规则在各种测验情境下的结果对比图

在图 2 中，根据是否采用随机缩减技术可以将 6 种规则分为两类。在所有的 12 种测验情境中，未采用随机缩减技术的 C-SPRT、P-SPRT、M-GLR 以及新提出的 Mahalanobis-SPRT 规则的 PCC 均较高，而采用随机缩减技术的 M-SCSPRT 以及新提出的 M-SCGLR 的 PCC 相对较低（但也都在 80%以上）。与此同时，采用随机缩减技术的 2 种规则的 ATL 明显低于未采用随机缩减的 4 种规则。也就是说，随机缩减的方法尽管可能损失一定的分类精度，但能较大幅度地缩短被试作答的测验长度。

考察本研究提出的两种新方法的表现。对于本文提出的 Mahalanobis-SPRT，在补偿性分界曲线的情境下，其总体上具有较高的 PCC：在题目间多维时，该方法的 PCC 仅仅略低于表现最好的 P-SPRT 方法，而在题目内多维时，该方法具有 6 种方法中最高的 PCC；而在非补偿性分界曲线的情境下，虽然该方法的 PCC 低于其他未使用随机缩减的方法，但是 ATL 也有相应的降低，而且可以看到其表现随能力维度间相关的升高有更大改善。对于本文提出的另一种终止规则（即 M-SCGLR），在几乎所有测验情境下，相比于同样采用随机缩减的

M-SCSPRT，它的 PCC 有较大提高，而 ATL 增加的却并不多。在使用非补偿性曲线和题目内多维的情境下，M-SCGLR 的 PCC 甚至能够接近未采用随机缩减技术的规则的水平。

考察能力维度间的相关水平对各终止规则的影响，可以发现：随着能力维度间相关系数ρ的增加，6 种终止规则的 ATL 有减少的趋势，而 PCC 则有升高的趋势。以 Mahalanobis-SPRT 规则为例，随着ρ的增加，其在每个ρ值下的四种测验情境里的平均 PCC 由 0.916 逐渐增加到 0.925 和 0.942，而平均 ATL 则由 57.037 下降到 54.437 和 52.384。考察分界曲线对各终止规则的影响，可以发现：相比于非补偿的分界曲线，6 种终止规则在几乎所有的补偿性分界曲线情境下的 ATL 均有所下降，而 PCC 则有所升高。考察题库结构对各终止规则的影响时，情况就变得复杂起来。由图 2 知，它与分界曲线会对各终止规则的表现产生交互作用。也就是说，在补偿性分界曲线的情境下，相比于题目间多维的条件，6 种终止规则在题目内多维条件下的 ATL 均有所下降，而 PCC 均有所升高。而在非补偿分界曲线的情境下，6 种规则在题目内多维与题目间多维的差异就没有统一规律。

图 3 呈现的是 6 种终止规则在各种 MCCT 测验情境下的标准化平均损失。图中的横坐标代表错误分类的惩罚R（详见公式（28）），其从区间[0,3000]按步长为 1 取值，共得到 3001 个点；纵坐标是在各个R值下 6 种规则的平均损失的标准化值[1]。

[1] 根据公式（28），随R值增大，平均损失也将不断增大。为清晰展示 6 种终止规则的平均损失值的相对关系随R的变化趋势，此处呈现在各个R值点处 6 种终止规则标准化后的平均损失。标准化后，我们只关注每一测验条件下 6 种规则的相对大小关系。

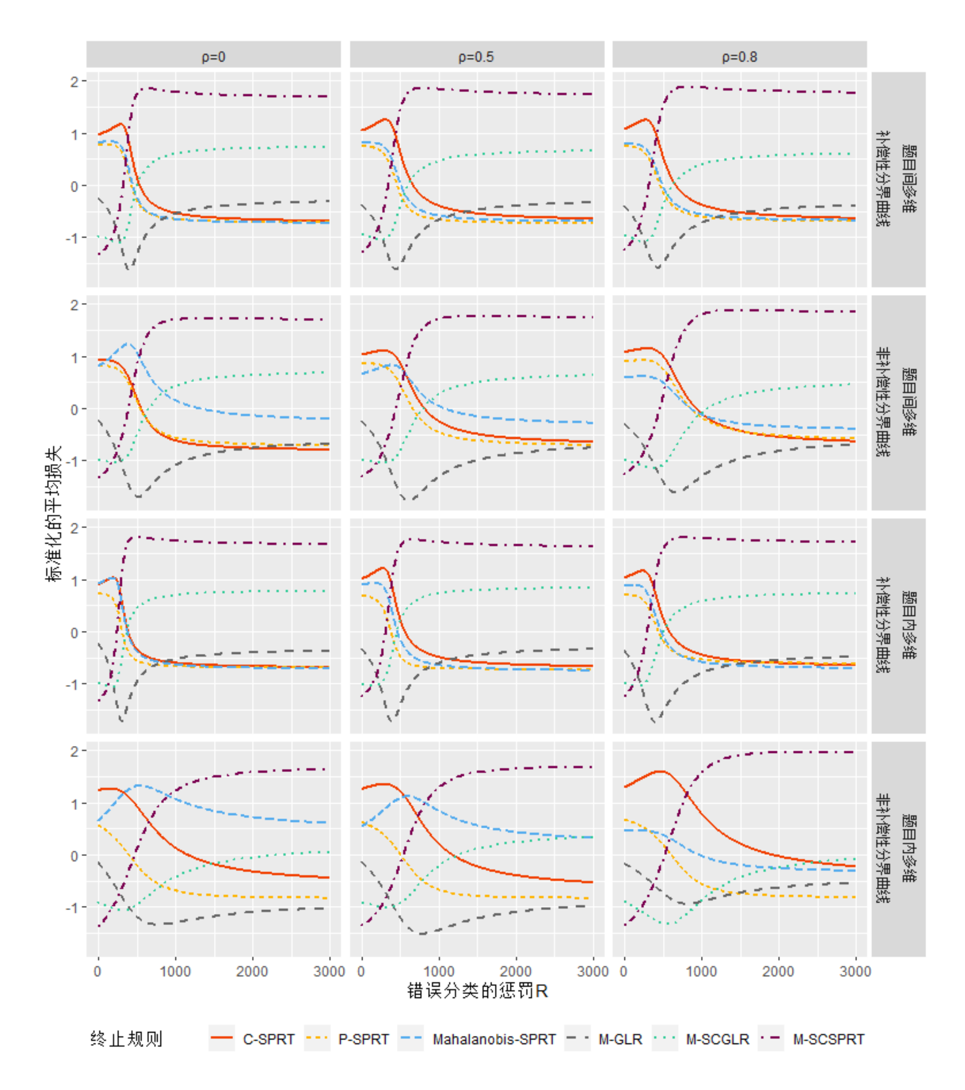


图 3 6 种终止规则在各种测验情境下的标准化平均损失变化图

根据平均损失的定义，随着R值的增加，平均损失对误判的敏感度不断上升。当R值约小于 500 时，ATL 较小的规则（即 M-GLR、M-SCGLR 和 M-SCSPRT）平均损失较小；当R值大于约 1000 时，平均损失对测验分类精度更敏感，分类精度较高的规则（即 C-SPRT、P-SPRT 和 Mahalanobis-SPRT）平均损失更小。

对于新提出的 Mahalanobis-SPRT，在补偿性分界曲线的所有情境下，该方法在对精度要求较高（即R取值较大）时，具有更低的平均损失（这与其在补偿性分界曲线的条件下具有更高的 PCC 是对应的）；在非补偿分界曲线的所有情境下，无论R的取值如何，其平均损失的值整体上处于 6 种方法的中间位置。这就是说，在实际测验中，该方法更适用于以下两种情境：一是使用补偿性分界曲线且对精度要求较高的情境。在该情境下，Mahalanobis-SPRT 的平均损失较其他规则更低，表现更好；二是使用非补偿性分界曲线且不能确定对精度的具体要求的情境。此时，虽然 Mahalanobis-SPRT 的平均损失并不是最低，但是由于测验对精度的要求并不确定，选择其他规则可能会导致在精度要求较高/较低时产生较大的损失。因

此，可以将 Mahalanobis-SPRT 作为一种相对“保守”的选择。对于本研究提出的另一种规则（即 M-SCGLR），在几乎所有测验情境下，该方法在R很小（约小于 200）时的平均损失略高于同样使用随机缩减技术的 M-SCSPRT。但是，在R的值稍高（约大于 200）时，M-SCGLR 的平均损失较 M-SCSPRT 有明显的降低（这与“其 PCC 较 M-SCSPRT 有较大提高，而 ATL 的增加则相对较少”相对应）。这表明在多数情境下，M-SCGLR 在测验精度上明显优于 M-SCSPRT。

4.2 各种规则对特定被试的敏感度

对应于第二个研究目的，图 4 和图 5 呈现了能力为各种特定值的被试在 6 种终止规则下的 PCC 结果。需要说明的是，图 4 中的黑色实线表示补偿的能力分界曲线$g(\boldsymbol{\theta}) = \theta_1 + \theta_2 = 0$，图 5 中的黑色实线则表示非补偿的能力分界曲线$g(\boldsymbol{\theta}) = \begin{cases} \theta_1 = 0, \theta_2 \geq 0 \\ \theta_2 = 0, \theta_1 > 0 \end{cases}$。

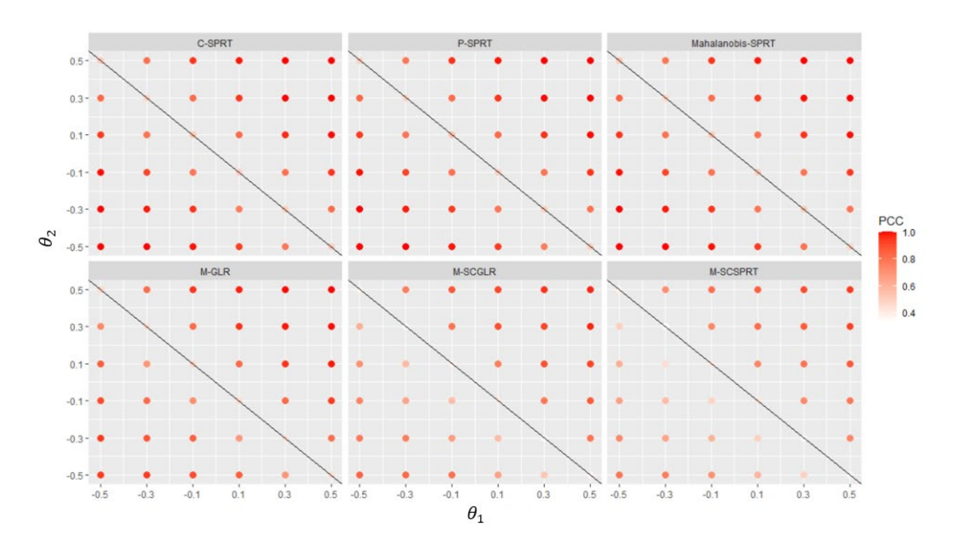


图 4 能力为各种特定值的被试在补偿性边界下 6 种终止规则的 PCC 结果

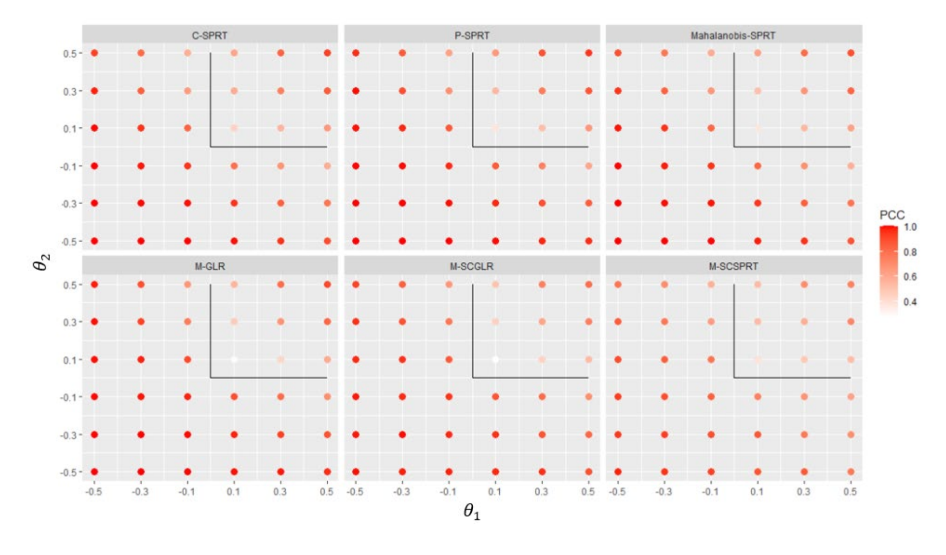


图 5 能力为各种特定值的被试在非补偿性边界下 6 种终止规则的 PCC 结果

由图 4 和图 5 可知，无论采用哪种分界曲线，6 种终止规则在 PCC 指标上对各种特定能力被试的敏感度呈现出一致的变化规律。具体而言，对于能力值靠近能力分界曲线的被试，其测验的 PCC 都较低；而对于能力值远离能力分界曲线的被试，其测验的 PCC 都较高。这说明能力值越靠近能力分界曲线的被试，越难对其进行准确的分类。上述能力为各种特定值的被试在 ATL 上呈现的规律则与 PCC 恰好相反。也即对于 6 种终止规则，能力值越靠近能力分界曲线的被试，其 ATL 越大。限于篇幅，此处不再呈现 ATL 的结果。

5 讨论及未来的研究方向

本研究采用测验分类精度及测验效率两个方面的指标，将新提出的两种 MCCT 终止规则与已有的终止规则进行比较。在单维 CCT 中，基于不同类别被试间的能力阈值即可构造似然比统计量，并进行假设检验，从而达到对被试进行分类的目的。但在 MCCT 中，由于不同类别被试的分界点变为分界曲线或曲面，故需要对传统 CCT 中的方法进行调整以适应这一变化。在已有的 MCCT 终止规则中，C-SPRT 与 P-SPRT 规则分别使用约束与投影的方式，将能力分界曲线或曲面“压缩”为分界点；M-GLR 规则对统计量的定义域进行一定的调整；M-SCSPRT 规则将随机缩减技术与 C-SPRT 相结合，大大提高测验效率。值得注意的是，由于 P-SPRT 仅使用一次估计的能力结果进行投影，在测验初期可能会使投影得到的$\hat{\boldsymbol{\theta}}_0$不够稳定，从而影响测验分类。对此，本文提出基于马氏距离的 Mahalanobis-SPRT 新规则，以弥补这一不足。另外，本文在 MCCT 情境中还对单维的 SCGLR 方法进行多维拓展，并得到 M-SCGLR 新规则。

根据 4.1 的结果，有一些值得讨论的发现：（1）对于本研究提出的 M-SCGLR 方法，在非补偿曲线和题目内多维的情境下，其 PCC 较高，接近未使用随机缩减技术的方法。这可能是因为以广义似然比为基础的 M-SCGLR 方法不需要“将分界曲线转化为分界点”，所以受分界曲线的影响更小，而其他规则的 PCC 在非补偿曲线的情境下会明显下降；（2）对于 Mahalanobis-SPRT 方法而言，尽管预期其能够弥补 P-SPRT 在测验前期所产生的能力估计问题，但是它在模拟结果中的表现并不尽如人意。这可能是由于该方法在测验后期时，会使用较多测验前期的作答信息，从而加大测验前期作答对结果的影响。本研究所设置的最大测验长度为 100，这意味着在测验结束时，往往能够得到比较准确的被试能力估计值。因此，Mahalanobis-SPRT 方法对 P-SPRT 初期的能力估计问题的弥补可能就无法很好体现。当最大测验长度较小时，该方法可能会有更好表现；（3）在未使用随机缩减的 4 种规则中，M-GLR 规则的 ATL 较其他 3 种规则有较大幅度的减少，这与 Thompson（2011）在单维情境下得到

的结论一致；（4）随着能力维度间相关系数ρ的增加，6 种终止规则都有更好的表现。这主要是因为增加维度间相关有助于提高能力向量的估计精度。这也与贝叶斯统计中的普遍观点（即从高度相关的维度中借用信息会产生更准确的能力估计）一致；（5）相比于非补偿的分界曲线，6 种终止规则在几乎所有的补偿性分界曲线情境下都有更好的表现，这是因为本研究所使用的 M3PL 模型是补偿性模型，其与补偿性的分界曲线更为契合，所以导致“在补偿性曲线情境下，各个终止规则的表现更好”；（6）题库结构与分界曲线会对终止规则的表现产生交互作用。具体而言，在补偿性分界曲线的情境下，相比于题目间多维的条件，6 种终止规则在题目内多维条件下的表现更好。这可能是因为相较于题目间多维的题库结构，题目内多维的题库结构能够提供更高的多维区分度，从而提高能力估计的准确性。具体来说，对于有着题目内多维结构的题库 1 来说，每个维度被所有 900 个题目测量；而对于有着题目间多维结构的题库 2 来说，每个维度只有 450 个题目测量（一半题目的$a_2 = 0$，另一半题目$a_1 = 0$）。但是在非补偿分界曲线的情境下，6 种规则在题目内多维与题目间多维的差异就没有统一规律。这可能是由于本研究考虑的非补偿边界其实就是直角坐标系中构成第一象限的坐标轴，所以边界上的能力阈值都只具有单一维度，导致题目内多维的上述优势不能很好发挥。

此外，还需要注意的一点是 MCCT 中的能力维度数量。理论上看，随着维度数的增加，平均测验长度会逐渐增加，而测验精度则会有下降趋势。但是，MCCT 是一个相当复杂的系统。当维度数不断增大时，平均测验长度和测验精度将会呈现何种变化趋势（是指数式的变化还是线性的变化？）、不同终止规则与不同选题策略的组合会对结果造成何种影响以及随机缩减技术的优势是否会进一步扩大，都有待进一步的研究。同时，当维度数增大到一定程度时，计算机还将会面临一些计算上的挑战。

本研究仍有一些不足之处，比如：本文主要局限于提出新规则以及模拟实现上，在对新规则理论性质的推导和证明等方面仍有待完善。本研究所讨论的规则均限定在对被试进行二分的情境下，而没有考虑多分类的情况。在模拟研究的设置上，本研究没有考虑非补偿 MIRT 模型、其他的多维区分度参数生成方式、不同的维度数以及不同的最大测验长度对结果的影响。

未来可以从以下四方面进一步开展研究：（1）提出新的多维似然比统计量。考虑构造将能力分界曲线或曲面转化为分界点的新方法，使得“在保证良好的分类准确率及测验效率的同时，能较好解决目前方法中存在的问题”；（2）开发多分类 MCCT 的终止规则。目前，有研究者对多分类的 CCT 终止规则进行探索（比如，Wang et al., 2020），但是对多分类 MCCT

终止规则的研究仍未见公开报道。构造多分类 MCCT 终止规则可实现在多维情境下对被试的更细致分类，值得今后进一步探索；(3)考虑融入过程性信息，比如反应时(response time)。已有研究表明，反应时能够提高能力估计精度（Wang & Hanson, 2005）。结合反应时构造 MCCT 终止规则，预期在保证测验效率的同时还能进一步提高分类精度。此外，由于马氏距离具有“不受量纲影响”和“能够同时考虑能力与反应时信息”等特点，因此本研究提出的 Mahalanobis-SPRT 可为这方面的探索提供一种可行路径；（4）在模拟研究中，考虑更丰富的条件设置（比如，不同的 MIRT 模型、多维区分度参数的生成方式、维度数以及最大测验长度等），考察其对结果的影响。

6 结论

模拟结果显示：（1）在使用补偿性分界函数的条件下，新提出的 Mahalanobis-SPRT 规则具有较高的分类精度以及与其他未使用随机缩减的方法相近的测验长度；（2）在几乎所有实验条件下，新提出的 M-SCGLR 规则不仅在测验精度大幅优于同样采用随机缩减的 M-SCSPRT 规则，而且具有较短的测验长度；（3）6 种终止规则在 PCC 和 ATL 上对具有不同能力被试的敏感度呈现出一致的变化规律。

致谢

感谢不列颠哥伦比亚大学（University of British Columbia）的在读博士生陈冠宇对本文摘要修改提供的帮助。

附录 1 M-SCGLR 中$P\left(D_J = D_{j'} \middle| C_{ij'}\right)$的推导

将被试作答至最大测验长度J时，对被试的分类判断与当前（作答j'道题）一致的概率记为$P\left(D_J = D_{j'} \middle| C_{ij'}\right)$。具体地说，如果目前的临时判断为被试属于“未掌握”类别，则被试作答至最大测验长度时仍被判断为“未掌握”的概率为$P\left(D_J = n \middle| C_{ij'}\right)$；如果目前的临时判断为被试属于“掌握”类别，则被试作答至最大测验长度时仍被判断为“掌握”的概率为$P\left(D_J = m \middle| C_{ij'}\right)$。不失一般性，我们对$P\left(D_J = n \middle| C_{ij'}\right)$的计算过程进行推导。

根据 Fineklman(2008)的研究，可以使用 Siegmund(1985)所描述的技巧。$P\left(D_J = n \middle| C_{ij'}\right)$实际上即为$P(C_{iJ} \le C_0 | C_{ij'})$。利用对数可加性以及中心极限定理，使用给定$C_{ij'}$下条件分布的渐近正态性，可以得到

$$P(D_J=n|C_{ij'})\approx\Phi\left(\frac{c_0-\mathbb{E}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})}{\sqrt{\mathrm{Var}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})}}\right). \tag{A1}$$

由于C_{iJ}中本身包含$C_{ij'}$的部分，根据对数的可加性及条件期望的性质，可得到公式（A1）中$\mathbb{E}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})$的计算，即

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})&=\mathbb{E}_{\boldsymbol{\theta}}\left(\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|\mathbf{Y}_{iJ})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|\mathbf{Y}_{iJ})]}\middle|C_{ij'}\right)\\&=\mathbb{E}_{\boldsymbol{\theta}}\left(\sum_{j=1}^{j'}\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}+\sum_{j=j'+1}^{J}\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}\middle|C_{ij'}\right)\\&=C_{ij'}+\sum_{j=j'+1}^{J}\mathbb{E}_{\boldsymbol{\theta}}\left(\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}\right).\end{aligned} \tag{A2}$$

类似地，根据作答反应的条件独立性及条件方差的性质，可得

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}}(C_{iJ}|C_{ij'})&=\mathrm{Var}_{\boldsymbol{\theta}}\left(\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|\mathbf{Y}_{iJ})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|\mathbf{Y}_{iJ})]}\middle|C_{ij'}\right)\\&=\mathrm{Var}_{\boldsymbol{\theta}}\left(\sum_{j=1}^{j'}\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}+\sum_{j=j'+1}^{J}\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}\middle|C_{ij'}\right)\\&=\sum_{j=j'+1}^{J}\mathrm{Var}_{\boldsymbol{\theta}}\left(\log\frac{\sup_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_m}[L(\boldsymbol{\theta}_1|Y_{ij})]}{\sup_{\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_n}[L(\boldsymbol{\theta}_2|Y_{ij})]}\right).\end{aligned} \tag{A3}$$

实际上，Huebner 和 Fina（2015）在单维情境下已对这一过程进行推导。与单维的 SCGLR 相比，在本文考虑的 MCCT 情境中，M-SCGLR 只是将$C_{ij'}$中求极值的集合由单维的区间变为多维空间中的区域，因此其推导过程是一致的。

附录 2 各终止规则的模拟结果

表 2 图 2 所对应的模拟结果

相关	分界曲线	题库结构	终止规则	PCC	ATL
			C-SPRT	0.948	52.959
			P-SPRT	0.948	49.541
$\rho=0$	补偿性	题目内多维	Mahalanobis-SPRT	0.950	53.216
			M-GLR	0.924	32.241
			M-SCGLR	0.858	18.849

			M-SCSPRT	0.807	12.649
		题目间多维	C-SPRT	0.930	61.981
			P-SPRT	0.929	57.835
			Mahalanobis-SPRT	0.930	58.876
			M-GLR	0.904	36.016
			M-SCGLR	0.851	20.848
			M-SCSPRT	0.805	13.504
	非补偿性	题目内多维	C-SPRT	0.908	69.070
			P-SPRT	0.915	55.622
			Mahalanobis-SPRT	0.873	57.369
			M-GLR	0.916	41.331
			M-SCGLR	0.879	26.151
			M-SCSPRT	0.829	17.048
		题目间多维	C-SPRT	0.931	61.163
			P-SPRT	0.927	58.847
			Mahalanobis-SPRT	0.909	58.686
			M-GLR	0.919	36.718
			M-SCGLR	0.864	20.974
			M-SCSPRT	0.825	14.012
$\rho = 0.5$	补偿性	题目内多维	C-SPRT	0.949	51.839
			P-SPRT	0.949	46.301
			Mahalanobis-SPRT	0.951	49.922
			M-GLR	0.929	28.306
			M-SCGLR	0.880	16.641
			M-SCSPRT	0.848	12.333
		题目间多维	C-SPRT	0.942	60.648
			P-SPRT	0.943	54.795
			Mahalanobis-SPRT	0.942	55.901
			M-GLR	0.921	32.052
			M-SCGLR	0.879	20.429
			M-SCSPRT	0.836	13.478
	非补偿性	题目内多维	C-SPRT	0.915	69.277
			P-SPRT	0.918	56.422
			Mahalanobis-SPRT	0.890	54.840
			M-GLR	0.917	41.205
			M-SCGLR	0.879	25.501
			M-SCSPRT	0.843	16.417
		题目间多维	C-SPRT	0.931	65.105
			P-SPRT	0.931	61.374
			Mahalanobis-SPRT	0.917	57.084
			M-GLR	0.925	37.549
			M-SCGLR	0.876	21.250
			M-SCSPRT	0.839	13.966

$\rho = 0.8$	补偿性	题目内多维	C-SPRT	0.960	50.987
			P-SPRT	0.957	45.382
			Mahalanobis-SPRT	0.961	48.457
			M-GLR	0.946	27.139
			M-SCGLR	0.896	16.513
			M-SCSPRT	0.858	12.313
		题目间多维	C-SPRT	0.958	58.903
			P-SPRT	0.958	52.540
			Mahalanobis-SPRT	0.958	53.414
			M-GLR	0.939	30.312
			M-SCGLR	0.897	19.343
			M-SCSPRT	0.851	13.860
	非补偿性	题目内多维	C-SPRT	0.920	68.485
			P-SPRT	0.928	56.274
			Mahalanobis-SPRT	0.916	52.433
			M-GLR	0.917	39.755
			M-SCGLR	0.902	25.742
			M-SCSPRT	0.856	16.835
		题目间多维	C-SPRT	0.944	65.928
			P-SPRT	0.941	61.900
			Mahalanobis-SPRT	0.933	55.232
			M-GLR	0.935	35.541
			M-SCGLR	0.898	20.446
			M-SCSPRT	0.857	14.111

参考文献

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18*, 257–275.

Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73*, 473–486.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.

Chen, P. (2016). Two new online calibration methods for computerized adaptive testing. *Acta Psychologica Sinica, 48*, 1184–1198.

[陈平. (2016). 两种新的计算机化自适应测验在线标定方法. *心理学报, 48*, 1184–1198.]

Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika, 81*, 674–701.

Chen, P., Wang, C., Xin T., & Chang, H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 70*, 81–117.

Finkelman, M. (2003). *An adaptation of stochastic curtailment to truncate Wald's SPRT in computerized adaptive testing* (CSE Report 606). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*, 442–463.

Finkelman, M. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*, 27–45.

Finkelman, M., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: a method to reduce respondent burden. *Statistics in Medicine, 30*, 1989–2004.

Guo, L., Zheng, C. J., & Bian, Y. F. (2015). Exposure control methods and termination rules in variable-length cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 47*, 129–140.

[郭磊，郑蝉金，边玉芳. (2015). 变长 CD-CAT 中的曝光控制与终止规则. *心理学报, 47*, 129–140.]

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Journal of Psychology, 216*, 89–101.

Huebner, A. R., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: a new termination criterion for variable-length computerized classification tests. *Behavior Research Methods, 47*, 549–561.

Kang, C., & Xin, T. (2010). New development in test theory: multidimensional item response theory. *Advances in Psychological Science, 18*, 530–536.

[康春花，辛涛. (2010). 测验理论的新发展:多维项目反应理论. *心理科学进展, 18*, 530–536.]

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.

Li, X., Zhang, J., & Chang, H. (2020). Look-ahead content balancing method in variable-length computerized classification testing. *British Journal of Mathematical and Statistical Psychology, 73*, 88–108.

Nydick, S. (2013). *Multidimensional mastery testing with CAT* (Unpublished doctoral dissertation). University of Minnesota.

Reckase, M. D., & McKinley, R. L. (1982). *Some latent trait theory in a multidimensional latent space*. Iowa City, IA: American College Service.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals.* Springer-Verlag.

Smits, N., & Finkelman, M. (2013). A comparison of computerized classification testing and computerized adaptive testing in clinical psychology. *Journal of Computerized Adaptive Testing, 1*, 19–37.

Thompson, N. A. (2010, June). *Nominal error rates in computerized classification testing*. Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, the Netherlands.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, & Evaluation, 16*, 1–7.

Wald, A. (1947). *Sequential analysis*. John Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *The Annals of Mathematical Statistics, 19*, 326–339.

Wang, C., Chen, P., & Huebner, A. (2020). Stopping rules for multi-category computerized classification testing. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.12202

Wang, T., & Hanson, B. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*, 323–339.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 450–480.

Two New Termination Rules for Multidimensional Computerized Classification Testing

REN He, CHEN Ping

(*Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing 100875, China*)

Abstract

Computerized classification testing (CCT) is a subset of computerized adaptive testing (CAT), and it aims to classify examinees into one of at least two possible categories that denote results such as pass/fail or non-mastery/partial mastery/mastery (Huebner & Fina, 2015). Therefore, CCTs focus on increasing the accuracy of classification which is different from CATs designed for precise measurement (Nydick, 2013). The termination rule is one of the key components of CCT. However, as pointed out by Nydick (2013), most CCTs (i.e., UCCTs) were designed under unidimensional item response theory (IRT), in which the unidimensionality assumption is easily violated in practice. Thus, researchers then began to construct multidimensional CCT termination rules (i.e., MCCT) based on multidimensional IRT. To date, however, these rules still have some deficiencies in terms of classification accuracy or test efficiency.

Most current studies on termination rules of MCCT are based on termination rules of UCCT. In UCCTs, termination rules require setting a cut point, θ_0, of the latent trait to calculate the statistics; and when they are extended from UCCT to MCCT, the cut point will become a classification bound curve or even a surface (i.e., $g(\boldsymbol{\theta}) = 0$). At this time, a question is how to convert the curve or surface into θ_0. To this end, the projected sequential probability ratio test (P-SPRT), constrained SPRT (C-SPRT; Nydick, 2013), and multidimensional generalized likelihood ratio (M-GLR) were respectively proposed to solve the problem in different ways. Among them, P-SPRT and C-SPRT choose specific points on g(θ) as the approximate cut point, $\hat{\theta}_0$, by projecting into Euclidean space or constraining on g(θ) respectively; as for M-GLR, because the generalized likelihood ratio statistic can be calculated without a cut point, it can be directly employed in MCCT. To overcome the limitation that P-SPRT may lead to unstable results at the beginning of the test, this study proposed the Mahalanobis distance-based SPRT (Mahalanobis-SPRT).

In addition, stochastic curtailment is a technique for shortening the test length by predicting whether the classification of participants will change as the test continues. This article also combined M-GLR with the stochastic curtailment and proposed M-GLR with stochastic curtailment (M-SCGLR).

A full-scale simulation study was conducted to (1) compare both the Mahalanobis-SPRT and M-SCGLR with the P-SPRT, C-SPRT, M-GLR, and multidimensional stochastically curtailed SPRT (M-SCSPRT) under varying conditions; (2) compare the classification performance of the above six termination rules for participants with specific abilities to explore whether there is a significant difference in the sensitivity of various rules to classify specific participants. To achieve the first research objective, three levels of correlation between dimensions ($\rho = 0$, 0.5, and 0.8), two item bank structures (within-item multidimensionality and between-item multidimensionality), and two kinds of classification boundary (compensatory boundary and non-compensatory boundary) were considered; to achieve the second objective, 36 specific ability points (θ_1, θ_2) were generated where $\theta_1, \theta_2 \in \{-0.5, -0.3, -0.1, 0.1, 0.3, 0.5\}$. The results showed that: (1) when the compensatory classification function was used, the Mahalanobis-SPRT

led to higher classification accuracy and similar test length to the rules without stochastic curtailment; (2) under almost all conditions, the M-SCGLR not only possessed higher precision but also maintained the short test length, compared to M-SCSPRT that also uses stochastic curtailment; (3) the six termination rules showed a consistent change in the sensitivity of the precision and test length to specific participants.

To sum up, two new MCCT termination rules (Mahalanobis-SPRT and M-SCGLR) are put forward in this article. Although the simulation results are very promising, several research directions merit further investigation, such as the development of MCCT termination rules for more than two categories, and the construction of MCCT termination rules by incorporating process data like the response time.

Key words computerized classification testing, termination rule, multidimensional item response theory, Mahalanobis distance, stochastic curtailment